Comparison of combination of dimensionality reduction and classification techniques for identifying tree species using integrated QuickBird imagery and Lidar data

Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Comparison of combination of dimensionality reduction and classification techniques for identifying tree species using integrated QuickBird imagery and Lidar data

Lien T. H. Pham
Lars Brabyn
Thanh Duc Dang
Henry Gouk

SPIE.

Lien T. H. Pham, Lars Brabyn, Thanh Duc Dang, Henry Gouk, "Comparison of combination of dimensionality reduction and classification techniques for identifying tree species using integrated QuickBird imagery and Lidar data," *J. Appl. Remote Sens.* **13**(1), 018502 (2019), doi: 10.1117/1.JRS.13.018502.

Comparison of combination of dimensionality reduction and classification techniques for identifying tree species using integrated QuickBird imagery and Lidar data

Lien T. H. Pham,[a,b,c,*] Lars Brabyn,[a] Thanh Duc Dang,[d] and Henry Gouk[e]

[a]The University of Waikato, Department of Geography and Environmental Planning, Hamilton, New Zealand

[b]Vietnam National University-Ho Chi Minh City, University of Science, Faculty of Environment, Ho Chi Minh City, Vietnam

[c]Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, Virginia, United States

[d]Singapore University of Technology and Design, Pillar of Engineering Systems and Design, Singapore

[e]The University of Waikato, Department of Computer Science, Hamilton, New Zealand

Abstract. Combining data of different types and from different sources for the classification of tree species has gained popularity recently, but training models on such datasets often requires more computational demands and does not always result in higher accuracy due to feature redundancy and irrelevance. Thus preprocessing data using dimensionality reduction (DR) methods can be employed to improve the classification accuracy and reduce computations. The objective of this research is to investigate and compare tree species classification performance for different classification algorithms [naive Bayes (NB), logistic regression (LR), random forest (RF), and support vector machine (SVM)], combined with various DR methods (correlation-based feature selection filter, information gain, wrapper methods, and principal component analysis). Two primary datasets are used—QuickBird and LiDAR, as well as derived topography data. When DR is used prior to classification, the NB classifier had a significant improvement in accuracy. SVM and RF had the best classification accuracy without DR. The overall accuracies (OA) of SVM and RF are 88.2% and 87.2% (kappa 0.84 and 0.83), respectively, followed closely by LR (OA: 84.8%, kappa: 0.79) and more distantly by NB (OA: 79%, kappa: 0.72). It is recommended to use SVM and RF without DR or NB with DR for tree species classification.
 [DOI: 10.1117/1.JRS.13.018502]

Keywords: tree species; dimensionality reduction; naive Bayes; logistic regression; random forest; support vector machine.

Paper 180616 received Jul. 22, 2018; accepted for publication Mar. 5, 2019; published online Mar. 20, 2019.

1 Introduction

Computer-aided techniques have improved the speed of classifying and mapping vegetation and other earth's surface objects viewed from remote sensing imagery.[1] The availability of various remote sensing images provides abundant information regarding not only spectral information but also texture data to classify floristic composition up to the species and association level.[2,3] More available data require more resources and efforts in object classification, but the results do not always achieve higher accuracy. Different supervised and unsupervised machine-learning techniques are helping to make the identification of objects easier than ever. However, more features (or variables) do not mean better classification results because of the existence of irrelevant and unuseful variables.[4] As a result, dimensionality reduction (DR) methods could be used to select suitable variables.

*Address all correspondence to Lien T. H. Pham, E-mail: phamlien24@gmail.com

1931-3195/2019/$25.00 © 2019 SPIE

Vegetation classification is among important applications of remote sensing in land use and land cover mapping and monitoring. Information on the spatial distribution of tree species, especially for those with high ecological, economic, and cultural significance as well as those under threat, is important for managers and policy makers when deciding on appropriate conservation strategies. Pohutukawa (*Metrosideros excelsa*) is an example; this is a native tree species in New Zealand which has been subject to fires and land clearance, and more recently possum (*Didelphimorphia*) browsing.[5] Pohutukawa is a multistemmed tree up to 25 m high with large, rounded crowns. Identifying and mapping this tree species have been mostly based on field data such as in Simpson,[6] which is costly and time-consuming to perform. Although the application of remote sensing has become widespread, so far, there has been only the research of Pham et al.[7] applying remote sensing technology for classifying these trees. It is also important to note that in Ref. 7, the researchers only employed a single machine-learning technique for DR [random forest (RF)] and vegetation species classification [support vector machine (SVM)].

Many studies have combined spectral information derived from multispectral or hyperspectral images with height information derived from LiDAR data, which has improved the accuracy of tree species classification.[7,8] These studies showed that information diversity from multiple datasets results in more features for classification. Multispectral data can provide spectral reflectance and texture features, which are useful information to distinguish different tree species/types. More specifically, spectral reflectance features are related to different characteristics of tree canopies including photosynthetic pigments of plant tissue, canopy structure such as branch density, and tree size.[9,10] Texture information is linked to foliage characteristics such as size and density and crown closure. In terms of LiDAR data, many features derived from LiDAR such as height features, return types, intensity, and relative crown base height features are related to vertical structure of individual trees; therefore, they are complementary to the multispectral data to improve the tree types/species classification. However, not all features are useful for classification and even unnecessary features may decrease the performance.

DR methods may be used to reduce the redundancy and irrelevance of features. However, studies on the comparison of various classifiers combined with DR methods for tree species classification is limited. There were studies which used single datasets and compared two machine-learning algorithms: RF and SVMs.[10,11]

This study attempted to fill the gap in understanding the combination between classifiers and DR techniques and compare the performance of a variety of machine-learning classification algorithms [naive Bayes (NB), logistic regression (LR), RF, and SVM], combined with different DR methods [correlation-based feature selection filter (CFS), information gain (InfoGain), wrapper methods (wrapper), and principal component analysis (PCA)]. We used the same dataset as in Ref. 7, which was a comprehensive set of data of field trips, Lidar, and spectral imagery for this experiment.

2 Materials and Methods

2.1 *Study Area and Datasets*

The research area was the eastern Coromandel region in New Zealand (36°48'30"S to 36°55'30"S latitude, and 175°38'30"E and 175°48'30"E longitude) [see Fig. 2(a)]. The site is characterized by different land cover types, but this research only focuses on identifying four important tree species/types: Pohutukawa (*Metrosideros excels*), Manuka/Kanuka (*Leptospermum scoparium/Kunzea ericoides*), other broadleaf species, and coniferous species. Popular coniferous species are rimu (*Dacrydium cupressinum*) and radiata pine (*Pinus radiate*). Common native broadleaved species are kohekohe (*Dysoxylum spectabile*), kowhai (*Sophora microphylla*), and puriri (*Vitex lucens*).[12]

A QuickBird image and a LiDAR point cloud were used due to the relationship between the spatial resolution of remote sensing imagery and area of tree canopies and data availability. The QuickBird image was captured on November 5, 2010 (Fig. 1) and has a panchromatic band (450 to 900 nm and 0.6-m spatial resolution) and four multispectral bands—blue (450 to 520 nm), green (520 to 600 nm), red (630 to 690 nm), and NIR (760 to 900 nm)—with 2.4-m spatial resolution. The QuickBird image was atmospherically corrected using ATCOR-3 developed

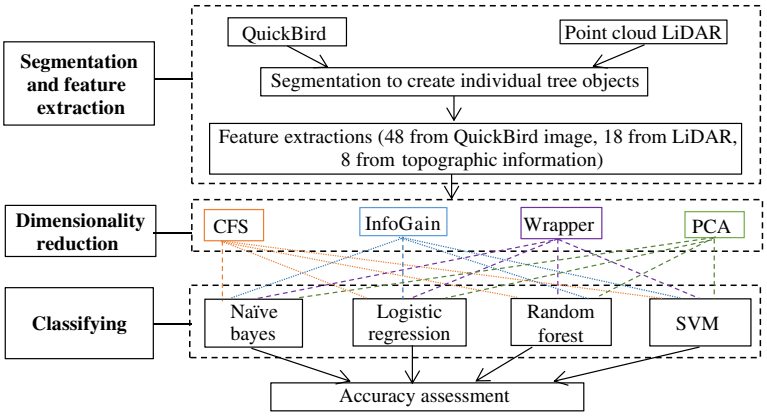

Fig. 1 Work flow of mapping tree species using DR and classification techniques.

by Ref. 13. The hue-saturation-intensity method was chosen for panchromatic sharpening to increase the spatial resolution to 0.6-m spatial resolution because this method can provide the best balance between the spectral and spatial information for QuickBird imagery when compared to other methods such as Gram–Schmidt and principal component.[14] The LiDAR dataset was captured during February and March 2013. The maximum number of returns for each pulse was four, and the average point density was 1.2 point/m^2.

For training and comparing the classification performance of feature selection and classification techniques, a ground-truth dataset was collected using a random sampling approach in November to December 2014. There were total 500 trees identified and distributed over the study area, 125 trees for each type of the four species/types in the region.

To evaluate the effects of DR methods on classifiers, a range of training set sizes, including 10, 25, 50, 75, 100, and 125 samples per class were used. This range is commonly used in remote sensing studies.[11,15] The supervised resampling filter in the WEKA Workbench[16] was used to generate small training sets from larger dataset containing ground truth information. The WEKA workbench includes advanced machine-learning algorithms and data preprocessing tools. It is designed to preprocess input data, evaluate machine-learning algorithms, and visualize the input data and the machine-learning results.

2.2 *Methods*

Identification of tree species using information derived from QuickBird and LiDAR data includes the following steps (an overview in Fig. 1). First, image segmentation, which divides an image into contiguous and separate areas called image objects (individual trees) was carried out automatically using eCognition. For each image object, there were seventy-four features extracted from QuickBird and LiDAR data, including four main categories: spectral, height, intensity, and topography (see in Table 1). The gray-level co-occurrence matrix (GLCM) and gray-level difference vector (GLDV) features were the texture features used. The relative height percentiles were calculated as the height percentile of laser returns divided by the maximum height of laser returns within individual tree crowns. The topographical wetness index used calculated using the formula developed by Ref. 17. Second step, DR was implemented to obtain a set of useful variables based on the 74 features. Then the image objects were classified into different species/types and were followed by the accuracy assessment to evaluate the performance of the different combined DR and classification methods. The performance of the classifiers with different levels of sample training data, ranging from 10 to 125 samples per class, was also compared.

The procedures used to segment individual trees and extract image features from the image objects are described in Ref. 7. For generating individual tree segments, first, object types such as grasslands or bare ground were masked using a threshold height value (lower than 2 m) derived from canopy height model (see Ref. 7 for more details). The remaining objects included trees and buildings. They were distinguished from each other by (1) using the multiresolution

Table 1 Image object features were used for classifications

Categories	Input layers	Object features	Number of features
Spectral	Blue	Mean of each layer	48
	Green	Standard deviation of each layer	
	Red	Texture variables of each layer: GLCM mean, GLCM standard deviation, GLCM correlation,	
	Nir	GLCM homogeneity, GLCM contrast, GLCM dissimilarity, GLCM entropy, GLDV mean, GLDV contrast, GLDV entropy	
Height	Point cloud LiDAR data	h_{mean1}: mean height of all returns within each tree crown	11
		h_{mean2}: mean height of first returns within each tree crown	
		h_{max}: maximum height of all returns within each tree crown	
		h_{min}: minimum height of all returns within each tree crown	
		reh_{10}: relative 10th height percentile of all returns within each tree crown	
		reh_{25}: relative 25th height percentile of all returns within each tree crown	
		reh_{50}: relative 50th height percentile of all returns within each tree crown	
		reh_{75}: relative 75th height percentile of all returns within each tree crown	
		reh_{90}: relative 90th height percentile of all returns within each tree crown	
		h_{st}: standard deviation of all returns within each tree crown	
		h_{coef}: coefficient of variation of all returns within each tree crown	
Intensity	Point cloud LiDAR data	i_{mean1}: mean intensity of all returns within each tree crown	7
		i_{mean2}: mean intensity of first returns within each tree crown	
		i_{max}: maximum intensity of all returns within each tree crown	
		i_{min}: minimum intensity of all returns within each tree crown	
		i_{st1}: standard deviation of all returns within each tree crown	
		i_{st2}: standard deviation of first returns within each tree crown	
		i_{coef}: coefficient of variation of all returns within each tree crown	
Topographic	DEM	Mean of each layer	8
	Slope	Standard deviation of each layer	
	Aspect		
	Topographic wetness index		

segmentation algorithm in eCognition to generate image objects with the four bands from the Quickbird image as input layers and the scale parameter set as 20; (2) classifying trees and buildings using values of elevation difference and normalized difference vegetation index. Trees can include a single-tree crown or a cluster of tree crowns. To delineate single-tree crowns from cluster ones, there are different algorithms that could be used, such as region growing, template matching, and valley-following. In this research, we used the region growing for delineating individual tree crown because it outperformed the other two methods.[18] The region growing method was run automatically using the image object fusion and parent process object

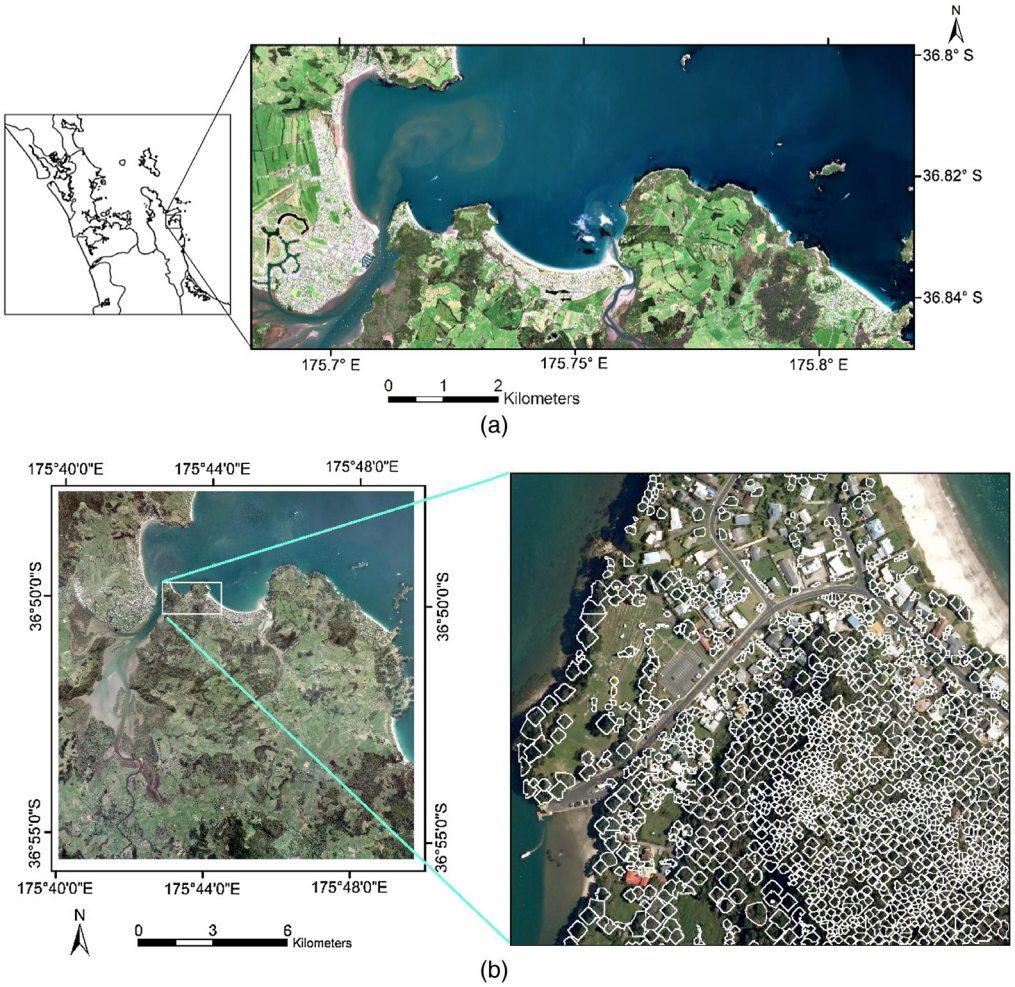

Fig. 2 (a) QuickBird image of the Coromandel study area: the coordinate is in NZTM2000 projection system and (b) individual tree crowns are represented by polygons.

in the eCognition software. These algorithms would merge treetop points with neighbor pixels that are similar and defined by crown extents, circular shapes, and the ration of length to width of the crown thresholds (see Ref. 7 about how to determine treetop points and criteria to define the values of thresholds). Figure 2(b) illustrates the individual crown trees obtained from the segmentation process. Features of individual trees including spectral and texture features such as Haralick's features, height features (mean height of laser returns, relative height percentiles of laser returns, etc.), intensity features (mean intensity of laser returns) were extracted using eCognition, while the topographic features were calculated using ArcGIS.

2.3 *Dimensionality Reduction Methods*

DR methods can be categorized as feature selection and feature extraction.[19]

2.3.1 *Feature selection methods*

The feature selection process selects a subset from the original feature set, including filter and wrapper methods.

Filter methods. Filter methods use feature ranking to select variables. A metric is used to compute the feature score, and then all features with a score below a user defined threshold are removed.[19] The advantages of filter methods are that they do not require high-computational

demand and operate independently of the choice of classifier.[20] Two filter methods were used in this research: correlation-based feature selection and InfoGain.

- CFS: CFS is a multivariate filter algorithm that uses a correlation-based heuristic evaluation function[21] to select a subset of features. These features are individually correlated with the class but uncorrelated with each other. The CFS's feature subset evaluation function is described by Ref. 22 as follows:

$$M_s = \frac{kr'_{cf}}{\sqrt{k + k(k-1)r'_{ff}}}, \tag{1}$$

  where M_s is the heuristic "merit" of a feature subset S containing k features, $\overline{r_{cf}}$ is the mean feature-class correlation ($f \epsilon S$), and $\overline{r_{ff}}$ is the average feature-feature intercorrelation.
- InfoGain: InfoGain is a metric for univariate filters which ranks features based on information value called entropy. The InfoGain value of a feature is a measure of the amount of uncertainty that is reduced for a target class when this feature is used. Features with higher InfoGain values have a greater probability of improving the classification.[22] As in Ref. 23, the formula for entropy and InfoGain are:

$$\text{Entropy} = \sum_{i=1}^{n}(-p_i \log_2 p_i) \text{ with } p_i \text{ is the probability of class } i, \tag{2}$$

$$\text{InfoGain}\,(\text{class}, \text{attribute}), \text{entropy}\,(\text{class}) - \text{entropy}\,(\text{class}|\text{attribute}). \tag{3}$$

Wrapper methods. Wrapper methods employ a search procedure to extract relevant feature sets and evaluate these feature sets using a machine-learning algorithm with cross validation. In the search procedure, the space of possible feature subsets is defined, and various subsets of features are generated. The evaluation of a specific feature subset is based on the classification accuracy of a classification model for this subset. Search algorithms can be broadly classified as sequential feature selection (SFS) and heuristic search.[19] The ability of interacting between feature subset search and classification model and taking feature dependencies into consideration are strengths of wrapper approaches. Forward selection and backward elimination are examples of the SFS. The SFS starts with no features (all features) and then sequentially add features (remove features) until the classification performance is not improved from the further addition (elimination).[23] The latter assesses different subsets which are created either by searching around in a search space or by creating solutions to the optimization problem.[19] Genetic algorithm and particle swarm optimization are two common algorithms of heuristic search. However, the disadvantages of using wrapper methods include having a higher possibility of over-fitting than filter methods and computationally expensive.[24]

In our research, the machine-learning algorithm used in the wrapper methods is the same for the classification. For example, the SVM used for classification process is also used in the wrapper methods. Although this can result in a lot of computation being required to build the classifier, it will prevent the feature selection process from over-fitting to a different inductive bias. We used linear forward selection as a search algorithm.

2.3.2 *Feature extraction method*

Feature extraction uses a transformation of the original features into a lower dimensional space. In this study, we focused on the PCA. PCA is one of the most popular feature extraction methods used in machine learning.[25,26] The first step is to find the principal components of the dataset. This is done by computing the covariance matrix of the data, and then performing an eigendecomposition. Once principal components have been obtained, a linear transformation can be constructed by concatenating the most informative eigenvectors into a matrix. By multiplying vectors from the original feature space by this matrix, the dimensionality of the data can be reduced while retaining most of the information.

Both feature selection methods (CFS, InfoGain, and wrapper methods) and feature extraction method (PCA) were implemented in the WEKA Workbench.

2.4 *Classification Techniques*

A range of classification algorithms were compared, including NB, LR, RF, and SVM.

2.4.1 *Naive Bayes classifier*

The NB classifier uses Bayes' theorem to predict a new instance and assumes that the predictive variables are independent given the output class.[27,28] The NB formula calculates the un-normalized posterior probabilities of each class using the following equation:

$$P(c|x_1, x_2, \ldots, x_n) = \frac{\prod_{i=1}^{n} P(x_i|c)P(c)}{P(x_1, x_2, \ldots, x_n)}, \tag{4}$$

where $P(\cdot)$ refers to the probability; $x_1, x_2, \ldots, x_n$ and are conditionally independent attributes given the class variable c. The instance is then categorized into the class associated with the highest un-normalized probability value.

2.4.2 *Logistic regression classifier*

LR classifiers are linear models for solving binary classification problems.[29] Because the data contained more than two classes, we selected one of the several generalizations to multiclass data. Namely, multinomial LR also goes by the name of SoftMax regression.[30] This method builds a set of binary LR models, each corresponding to a different class. Each of these models can then be used to compute a score indicating how likely a novel instance is to belong to each class. By normalizing this vector of scores, we can produce a categorical distribution over the possible classes. To make a prediction, one simply selects the class corresponding to the highest probability. Like building LR models, these SoftMax regression models can be trained by minimizing the negative log-likelihood of the model parameters using numerical optimization algorithms.

2.4.3 *Random Forest classifier*

An RF is an ensemble method that combines multiple decision trees by aggregating their predictions and treating them as votes.[31] Each tree is built from a bootstrap sample generated by sampling data randomly with replacement from the original dataset. A random subset of the features is used when determining the best split at each node of the tree—a technique known as feature bagging.[32,33] New instances are classified based on the majority vote of the decision trees in the ensemble. Two parameters—m_{try} (the number of predictors) and n_{tree} (the number of classification trees—need to be specified. These two parameters were identified using the multisearch scheme in the WEKA Workbench.

2.4.4 *Support Vector Machines classifier*

SVM algorithm, formally developed by Vapnik,[34] implicitly maps the original training data into a higher dimensional vector space, through the use of a kernel function, where the maximum margin separating hyperplane is used to classify the input data.[35,36] In this research, the Gaussian radial basis function (RBF) was used as the kernel function. Given the training data T with n samples: $T = \{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\}$, $x_i \in R^d$ and $y_i \in \{-1, 1\}$, $i = 1,2, \ldots, n$, the SVM classification model can be written as

$$f(x) = \text{sign}\left[\sum_{i,j=1}^{n} \alpha_i y_i K(x_i, x_j) + b\right], \tag{5}$$

where K is the kernel function, α_i is a Lagrange multiplier, and b is a scalar bias term.

The RBF-SVM was used for many reasons. First, the RBF kernel can deal with the situation where the relationship between the class labels and features is nonlinear. Second, the RBF kernel has fewer tuning parameters than the polynomial and the sigmoid kernels. There are two parameters required for an RBF-SVM: the regularization coefficient (C) and the kernel smoothness (gamma). The multisearch function in WEKA Workbench was used to find a good assignment for these two parameters. Values in the set 2^i with $i = -10, -9, \ldots, 15$ were considered for C, and 2^i with $i = -8, -2, \ldots, 8$ for gamma.

2.5 *Validation and Comparison Method*

To obtain a reliable result for each algorithm, a 10-fold cross validation was performed on the entire data set and repeated 5 times. For each 10-fold cross-validation process, the dataset was first divided into 10 equal-sized parts or folds. Subsequently, 10 iterations of training and validation were performed. With each iteration, a different fold was held out for validation and the remaining nine folds were used for training the classification model.

The overall accuracy (OA) and the kappa coefficient of agreement were used to measure the accuracy of each classifier with and without using DR. In addition, the McNemar's test, a nonparametric test was performed to compare different classifiers as suggested by Foody,[37] Pal and Foody.[11] The McNemar's test is based on the following equation:

$$z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}}, \tag{6}$$

where f_{12} and f_{21}, respectively, are the number of objects correctly identified by one classifier and not the other.

3 Results and Discussions

DR is often employed to reduce the number of bands for hyperspectral data products,[38] but in this study, we focused on using DR in reducing irrelevant spectral and texture variables from multispectral data and other features from LiDAR data such as height features. The DR techniques that we used had the same principle as the DR were used for hyperspectral data, which aimed to remove noise and redundant variables. Nevertheless, the latter technique has not been assessed for different combinations of DR and classifiers for vegetation classification yet.

This research compares four different classifiers combined with four different DR methods, and six different levels of sampling; thus, a total of 96 unique combinations are analyzed producing many results. The results include comparing (1) the performance of the four different classifiers combined with the four different DR methods, (2) the performance of the different classifiers using a range of sample sizes and no DR, and (3) the best combinations of classifier and DR (or no DR) are presented using different sample sizes. Figure 3 illustrates tree classification using SVM without DR.

The accuracy of the NB classifier was improved using DR method, but not RF and SVM. Table 2 and Fig. 4 show the performance of each classifier with 125 training samples per class combined with four different DR methods and no DR with the McNemar test results. The CFS and wrapper methods improved the accuracy significantly for the NB classifier. The NB is a probabilistic classifier based on the Bayes theorem, in which each feature NB_i is conditionally and ideally independent from every other features NB_j ($i \neq j$).[27] This means that NB is sensitive to redundant and irrelevant features, and this confirmed what was found in Ref. 38. The application of DR in this case, thus, improved the performance of classification. DR techniques did not improve the performance of RF, and when RF was used with PCA, it reduced the accuracy. RF performs an implicit feature selection or feature weighting in its learning process;[31] therefore, reducing the number of input features does not enhance the classification performance. The transformation of original variables into principal components (in PCA) may not necessarily retain all the information in the original dataset so the accuracy of combined PCA and RF was reduced. In the case of SVM, the input data can be grouped using different hyperplanes.[35,36] DR would help reduce redundant data, but it also weakens the prediction

Fig. 3 Tree objects were classified by the SVM without DR.

Table 2 Comparison between four different classifiers combined with four different DR methods and no DR with 10-fold CV repeated 5 times and McNemar's test at the 95% confidence level ($Z \geq |1.96|$).

Classifiers	DR method	OA (%)
NB	CFS	77.7 (+)
	InfoGain	72.9 (*)
	PCA	75.7 (*)
	Wrapper NB + linear forward selection	79.0 (+)
	No DR	71
LR	CFS	83.5 (*)
	InfoGain	83.9 (*)
	PCA	81.6 (*)
	Wrapper LR + linear forward selection	84.1 (*)
	No DR	84.8
RF	CFS	85.3 (*)
	InfoGain	86.2 (*)
	PCA	78.6 (-)
	Wrapper RF + linear forward selection	87.0 (*)
	No DR	87.2
SVM	CFS	85.9 (*)
	InfoGain	85.8 (*)
	PCA	84.9 (-)
	Wrapper SVM + linear forward selection	87.0 (*)
	No DR	88.2

Note: (+) and (-) denote that the result was statistically better or worse, respectively, than the no DR, whereas (*) denotes that there was no significant difference.

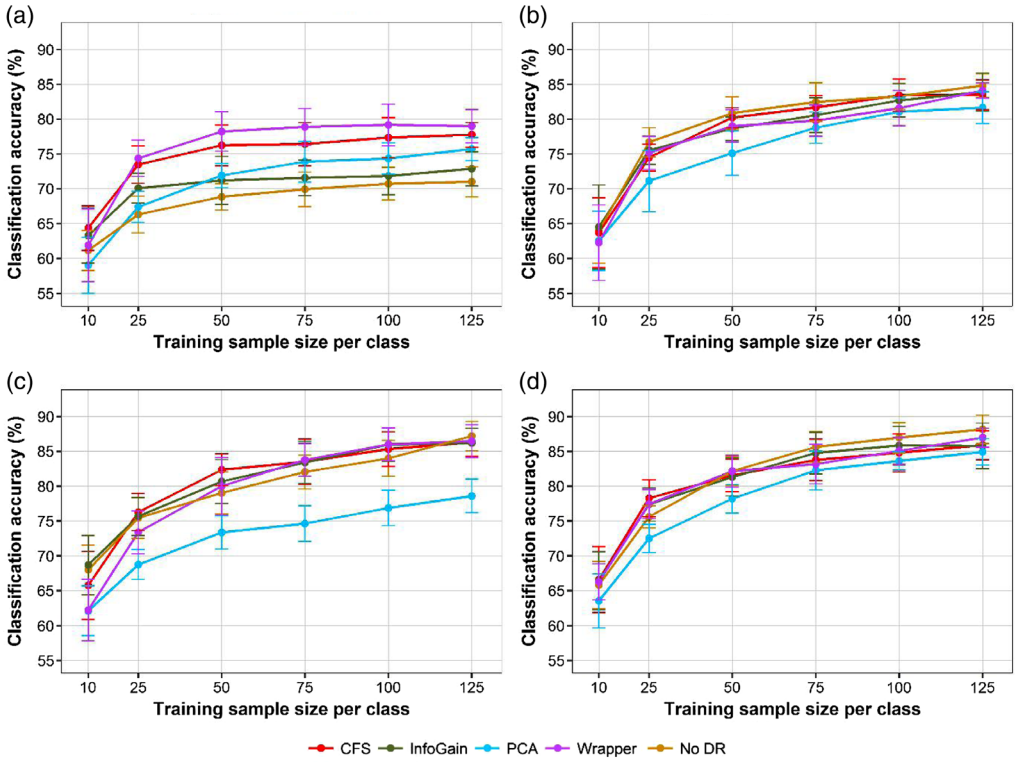

Fig. 4 Classification accuracy (OA) of (a)–(d) NB, LR, RF, and SVM with different DR and no DR using a range of training samples per class, respectively.

about the location of the hyperplanes. Since DR also did not improve most of the classifiers' performance, we can conclude that the more dimensions of the data are considered, the smaller error rates globally are. In theory, classifying techniques combined with DR should reduce noise in collected data, but since the original dimension of remote sensing data is not high, the advantage of DR is not clear.

We also notice that the performance of the classifiers became better when the number of samples increased. Figure 4 also shows the effect of training sample size on classification performance. Figure 5, for visual convenience, only compares the classifiers without DR. The accuracy of the NB classifier increases much slower than the other methods once the sample sizes are over 50, or the NB classifier works well with small training datasets. This confirms what was found in other 15 machine-learning applications in Ng and Jordan.[39] Table 3 supports these finding by showing the OA and kappa values with the level of statistical significance.

Each combination between classifiers and DR performed differently with a sample size. Table 4 and Fig. 6 show the comparison in the performance of each classifier when combined with the best DR (or no DR) in the previous experiments that suits each classifier for a particular training sample size. When the training sample size is 25 or less there is no difference in the performance of the different classifiers when these are combined with the most suitable DR (or no DR). This is because when a small sample size was used, it could result in nonsignificant results of the tests for the difference among classifiers.[36] When the sample size is 50 or greater, SVM is better than NB. SVM also has higher performance than RF when sample size is large, but this is not significant. RF is better than SVM when the sample size is 10 but this is also not significant. RFs and SVMs are both capable of approximating arbitrary nonlinear functions, with the accuracy of this approximation constrained by the regularization parameters of each scheme. For LR, although in general its OA and kappa were lower than RF, no statistically significant difference was found. This trend does not extend to the comparison with SVM, where LR was statistically significantly worse than SVM when the sample size is 125. The ability of these nonlinear methods to outperform LR provides some evidence that there is an inherent nonlinear relationship between the input features and the log odds of the output target variable.

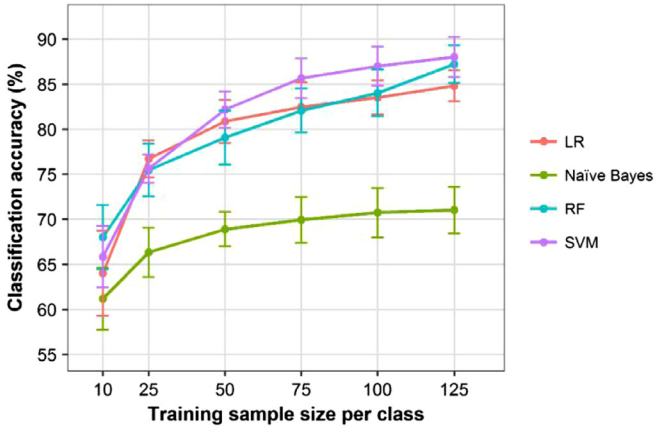

Fig. 5 Classification accuracy (OA) of NB, LR, RF, and SVM with different training sizes per class with no DR.

Table 3 Comparing NB, LR, RF, and SVM using all features and different training set sizes with 10-fold CV repeated 5 times and McNemar's test at the 95% confidence level ($Z \geq |1.96|$).

Training size (samples per class)	Classifier	OA (%)	Kappa	McNemar's test			
				NB	LR	RF	SVM
10	NB	61.2	0.48	NB	*	*	*
	LR	64.0	0.52	*	LR	*	*
	RF	**68.0**	**0.57**	+	*	RF	*
	SVM	65.8	0.54	*	*	*	SVM
25	NB	67.3	0.56	NB	-	-	-
	LR	**76.7**	**0.69**	+	LR	*	*
	RF	75.4	0.67	+	*	RF	*
	SVM	75.6	0.67	+	*	*	SVM
50	NB	67.9	0.57	NB	-	-	-
	LR	80.8	0.74	+	LR	*	*
	RF	79.0	0.72	+	*	RF	*
	SVM	**82.2**	**0.76**	+	*	*	SVM
75	NB	69.9	0.60	NB	-	-	-
	LR	82.4	0.77	+	LR	*	*
	RF	82.0	0.76	+	*	RF	*
	SVM	**85.6**	**0.81**	+	*	*	SVM
100	NB	70.7	0.61	NB	*	*	*
	LR	83.3	0.78	+	LR	*	-
	RF	84.0	0.79	+	*	RF	*
	SVM	**87.0**	**0.83**	+	+	*	SVM
125	NB	71.0	0.61	NB	-	-	-
	LR	84.8	0.79	+	LR	*	-
	RF	87.2	0.83	+	*	RF	*
	SVM	**88.2**	**0.84**	+	+	*	SVM

Note: Bolded OA and kappa scores indicate the highest performance; + and - denote that the result is statistically better or worse than the classifier compared; whereas * denotes there is no significant difference between classifiers.

Table 4 Comparing the best classification accuracy of NB, LR, RF, and SVM and different training set sizes with 10-fold CV repeated 5 times and McNemar's test at the 95% confidence level ($Z \geq |1.96|$)

Training size (samples)	Classifier + DR (or no DR)	OA (%)	Kappa	McNemar's test			
				NB	LR	RF	SVM
10	NB + CFS	64.4	0.52	NB	*	*	*
	LR + InfoGain	64.5	0.53	*	LR	*	*
	RF + InfoGain	**68.7**	**0.58**	*	*	RF	*
	SVM + CFS	66.6	0.56	*	*	*	SVM
25	NB + wrapper	74.3	0.66	NB	*	*	*
	LR + no DR	76.7	0.69	*	LR	*	*
	RF + CFS	76.3	0.68	*	*	RF	*
	SVM +CFS	**77.4**	**0.70**	*	*	*	SVM
50	NB + wrapper	77.2	0.71	NB	*	-	-
	LR + no DR	80.8	0.74	*	LR	*	*
	RF + CFS	**82.4**	**0.76**	+	*	RF	*
	SVM + no DR	**82.4**	**0.76**	+	*	*	SVM
75	NB + wrapper	77.7	0.71	NB	*	-	-
	LR + no DR	82.4	0.77	*	LR	*	*
	RF + CFS	83.8	0.78	+	*	RF	*
	SVM + no DR	**85.6**	**0.81**	+	*	*	SVM
100	NB + wrapper	78.3	0.72	NB	-	-	-
	LR + CFS	83.6	0.78	+	LR	*	*
	RF + wrapper	86.7	0.81	+	*	RF	*
	SVM + no DR	**87.0**	**0.83**	+	*	*	SVM
125	NB + wrapper	79.0	0.72	NB	-	-	-
	LR + no DR	84.8	0.79	+	LR	*	-
	RF + no DR	87.2	0.83	+	*	RF	*
	SVM + no DR	**88.2**	**0.84**	+	+	*	SVM

Note: Bolded OA and kappa scores indicate the highest performance; + and - denote that the result is statistically significantly better or worse, respectively, than the classifier compared; whereas * denotes there is no significant difference between classifiers.

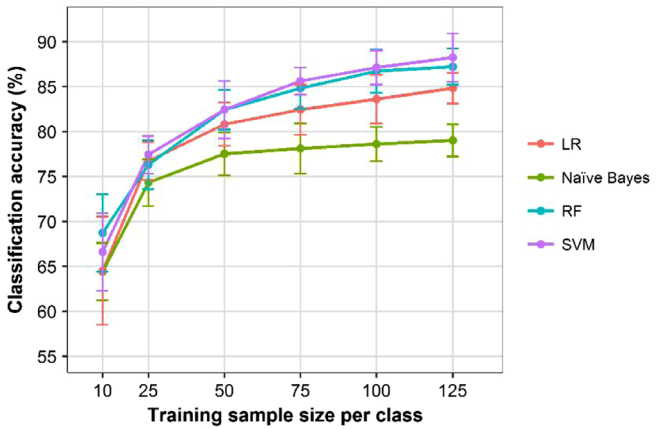

Fig. 6 The classification accuracy of the best combination of classifier and DR (or no DR) for different training sample sizes per class.

This study, however, contains some uncertainties. Optical data were acquired in 2010, LiDAR data were acquired in 2013, and the field data were collected in November to December 2014. Although the temporal differences between these datasets, there was a little impact on the classification for the following reasons. The differences between years among the three datasets were not large and the classified trees were large trees (higher than 2 m). As a result, there was not much difference between vertical structures of LiDAR with the field data and between spectral information from Quickbird data with the field data. Second, the QuickBird images were orthorectified by the vendor to with RMS error of 2.6 m.[40] The RMSE value is higher than the spatial resolution of the pansharpened image. When the RMSE values of images "overlaid" the GPS locations of the trees, it can reduce identifying of the trees. However, the RMSE values are quite small compared to the tree crown in this study (crowns are larger than 4 m and height are higher than 2 m). Therefore, it does not impact much the capacity of identifying trees in this study.

4 Conclusion

A range of classifiers and DR methods has been systematically studied in this research for tree species classification to provide a clear picture of how these techniques perform and the trade-offs associated with different sampling intensities. The four DR methods did not improve the accuracy of SVM and RF, but improved the accuracy of NB. This research does not show that DR methods significantly decrease the accuracy of classifiers; therefore, they can be applied prior to classification with all four classifiers to accelerate the model fitting process—particularly when the sample size is large.

This study has provided further evidence that SVM and RF are the best classification algorithms compared to other machine-learning algorithms when using integrated datasets for identifying tree species. Although NB was not as good as others in terms of classification accuracy, when the sample size was small, it achieved similar accuracies to the other classifiers. This is because generative models tend to outperform discriminative models when sample sizes are small. The additional benefit of NB is that it is simple and fast compared to LR, RF, and SVM because it does not require tuning parameters for training.

Future research could investigate the performance of other generative model, such as Bayesian Networks, that could potentially be more sample efficient than discriminative models such as LR, RF, and SVM.

Acknowledgments

Lien Pham would like to thank the New Zealand Aid Program for providing financial assistance to pursue this research. The authors greatly appreciate three anonymous reviewers for their thoughtful comments and suggestions. Declaration of interest statement: none.

References

1. P. M. Mather and M. Koch, *Computer Processing of Remotely-Sensed Images: An Introduction*, John Wiley & Sons, West Sussex, Chichester (2011).
2. U. Nguyen et al., "Mapping vegetation types in semi-arid riparian regions using random forest and object-based image approach: a case study of the Colorado River Ecosystem, Grand Canyon, Arizona," *Ecol. Inf.* **50**, 43–50 (2019).
3. L. T. Pham et al., "Monitoring mangrove association changes in the Can Gio Biosphere Reserve and implications for management," *Remote Sens. Appl. Soc. Environ.* **13**, 298–305 (2019).
4. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1226–1238 (2005).
5. R. Bylsma, B. Clarkson, and J. Efford, "Biological flora of New Zealand 14: Metrosideros excelsa, Pohutukawa, New Zealand Christmas tree," *N. Z. J. Bot.* **52**, 365–385 (2014).

6. P. G. Simpson, *Natural Pohutukawa in Taranaki*, Department of Conservation, New Zealand (1998).
7. L. T. Pham, L. Brabyn, and S. Ashraf, "Combining QuickBird, LiDAR, and GIS topography indices to identify a single native tree species in a complex landscape using an object-based classification approach," *Int. J. Appl. Earth Obs. Geoinf.* **50**, 187–197 (2016).
8. Y. Ke, L. J. Quackenbush, and J. Im, "Synergistic use of QuickBird multispectral imagery and LIDAR data for object-based forest species classification," *Remote Sens. Environ.* **114**, 1141–1154 (2010).
9. F. E. Fassnacht et al., "Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass," *Remote Sens. Environ.* **154**, 102–114 (2014a).
10. F. E. Fassnacht et al., "Comparison of feature reduction algorithms for classifying tree species with hyperspectral data on three central European test sites," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.* **7**, 2547–2561 (2014b).
11. M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.* **48**, 2297–2307 (2010).
12. E. A. Humphreys and A. M. Tyler, *Coromandel Ecological Region: Survey Report for the Protected Natural Areas Programme*, Department of Conservation, Waikato Conservancy, New Zealand (1990).
13. R. Richter and D. Schläpfer, *ATCOR-2/3 User Guide, Version 8.3. 1*, ReSe Applications Schläpfer, Wil, Switzerland (2014).
14. Y. Zhang and R. K. Mishra, "A review and comparison of commercially available pan-sharpening techniques for high resolution satellite image fusion," in *IEEE Int. Geosci. and Remote Sens. Symp.*, pp. 182–185 (2012).
15. L. Ma et al., "Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers," *ISPRS Int. J. Geo-Inf.* **6**, 51 (2017).
16. E. Frank, M. Hall, and I. Witten, *The WEKA Workbench*, 4th ed., Morgan Kaufmann, Burlington, Massachusetts (2016).
17. K. J. Beven and M. J. Kirkby, "A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant," *Hydrol. Sci. J.* **24**(1), 43–69 (1979).
18. M. Larsen et al., "Comparison of six individual tree crown detection algorithms evaluated under varying forest conditions," *Int. J. Remote Sens.* **32**(20), 5827–5852 (2011).
19. G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.* **40**(1), 16–28 (2014).
20. S. Galelli et al., "An evaluation framework for input variable selection algorithms for environmental data-driven models," *Environ. Modell. Software* **62**, 33–51 (2014).
21. M. A. Hall, "Correlation-based feature selection for machine learning," Doctoral Dissertation, The University of Waikato, Hamilton, New Zealand (1999).
22. B. Rogers et al., "Using text mining techniques to extract rationale from existing documentation," in *Design Computing and Cognition '14*, J. S. Gero and S. Hanna, Eds., Springer International Publishing, Cham, pp. 457–474 (2015).
23. I. H. Witten, E. Frank, and A. H. Mark, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Burlington, Massachusetts (2011).
24. Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics* **23**(19), 2507–2517 (2007).
25. T. D. Dang, T. A. Cochrane, and M. E. Arias, "Quantifying suspended sediment dynamics in mega deltas using remote sensing data: a case study of the Mekong floodplains," *Int. J. Appl. Earth Obs. Geoinf.* **68**, 105–115 (2018).
26. H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Syst.* **24**, 1024–1032 (2011).
27. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York (1973).

28. D. Soria et al., "A 'non-parametric' version of the naive Bayes classifier," *Knowledge-Based Syst.* **24**, 775–784 (2011).
29. D. R. Cox, "The regression analysis of binary sequences," *J. R. Stat. Soc. Ser. B (Methodol.)* **20**, 215–242 (1958).
30. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York (2007).
31. L. Breiman, "Random forests," *Mach. Learn.* **45**, 5–32 (2001).
32. A. Liaw and M. Wiener, "Classification and regression by random forest," *R News* **2**, 18–22 (2002).
33. V. F. Rodriguez-Galiano et al., "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.* **67**, 93–104 (2012).
34. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York (1995).
35. J. Han, M. Kamber, and J. Pei, *9 - Classification: Advanced Methods, Data Mining*, 3rd ed., Morgan Kaufmann, Boston, pp. 393–442 (2012).
36. G. P. Petropoulos, C. Kalaitzidis, and K. Prasad Vadrevu, "Support vector machines and object-based classification for obtaining land-use/cover cartography from Hyperion hyperspectral imagery," *Comput. Geosci.* **41**, 99–107 (2012).
37. G. M. Foody, "Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority," *Remote Sens. Environ.* **113**(8), 1658–1663 (2009).
38. J. Khodr and R. Younes, "Dimensionality reduction on hyperspectral images: a comparative review based on artificial data," in *4th IEEE Int. Congress Image and Signal Process. (CISP)*, Vol. 4, pp. 1875–1883 (2011).
39. A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes," in *Proc. 14th Int. Conf. Neural Inf. Process. Syst. Natural and Synthetic*, Vol. 2, pp. 841–848 (2002).
40. Land Information New Zealand, "KiwImage Products," 2014, https://www.linz.govt.nz/land/maps/linz-topographic-maps/imagery-orthophotos/kiwimage/kiwimage-products (15 January 2017).

Lien T. H. Pham received her PhD degree in geography in 2018 from the University of Waikato, New Zealand. Currently, she is a postdoctoral research fellow at the Virginia Institute of Marine Science, William and Mary, United States, where she is involved in mapping submerged aquatic vegetation. Her current research interests include remote sensing image processing, machine learning algorithms, and their large-scale applications.

Lars Brabyn is a senior lecturer in the geography program at the University of Waikato, specializing in GIS and remote sensing. The remote sensing applications he has helped develop include monitoring water quality, assessing the impacts of tourism on land-cover in Nepal, the use of night lights as an economic indicator, and using MODIS thermal IR data to produce a melt rate model for Antarctic glaciers.

Thanh Duc Dang is currently a postdoctoral research fellow at Resilient Water System Group, Pillar of Engineering Systems and Design, Singapore University of Technology and Design. His research interests include remote sensing image processing and water resources modeling.

Henry Gouk is working toward a PhD with the Department of Computer Science, the University of Waikato, Hamilton, New Zealand.